Final Project Technology Review

CS 410: Text Information Systems, Fall 2022
Kevin Eveker, keveker2@illinois.edu

Introduction

The topic that I selected for my technology review is Language Models: [Approach] OpenAI GPT, GPT-2, GPT-3 and BERT. I chose this topic in order to gain a better understanding of these tools and to compare and contrast their respective approaches in order to prep for my final project that will make use of GPT-3 to assist with better IMDB searches.

GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) are deep learning models that "process sequential input data, such as natural language, with applications towards tasks such as translation and text summarization (see https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)). These transformers differ from recurrent neural networks (RNNs) in that the process the text all at one time as opposed to one word at a time.

Transformer Language Models

OpenAI GPT

GPT is a massive neural net trained using internet data to generate text that seems like a human could have created it. GPT is an acronym for Generative Pre-trained Transformer, and GPT-3 (https://github.com/openai/gpt-3) is the 3rd generation of this model (see https://www.techtarget.com/searchenterpriseai/definition/GPT-3 for a nice high level overview and video). GPT can be tailored to a specific topic by using a topic specific data set. According to https://arxiv.org/pdf/2005.14165.pdf, GPT-3 does well in certain situations without tailoring:

- translation
- question-answering
- cloze tasks (filling in the blanks)
- on-the-fly reasoning or domain adaptation, such as:
    - unscrambling words
    - using a novel word in a sentence
    - performing 3-digit arithmetic.

This paper also states that the primary limitation to the approach taken by GPT developers is that "there is still a need for task-specific datasets and task-specific fine-tuning: to achieve strong performance on a desired task typically requires fine-tuning on a dataset of thousands to hundreds of thousands of examples specific to that task." The good news is (especially for our planned final project), OpenAI provides an easy to use API to tailor GPT-3 for a specific topic (see https://openai.com/blog/customized-gpt-3/). For GPT-3 (https://en.wikipedia.org/wiki/GPT-3), the model contains 175 billion parameters, compared to 1.5 billion parameters for GPT-2 (https://en.wikipedia.org/wiki/GPT-2) and 117 million parameters in GPT-1 (https://360digitmg.com/types-of-gpt-in-artificial-intelligence)
https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

BERT

BERT is an acronym for Bidirectional Encoder Representations from Transformers and is a pretrained transformer based NLP model developed by Google and use as part of their search engine (ses https://en.wikipedia.org/wiki/BERT_(language_model)). Like GPT, BERT can also be fine-tuned for specific tasks using a smaller task specific data set (see https://github.com/google-research/bert/blob/master/README.md#fine-tuning-with-bert). BERT code and pre-trained models can be found at https://github.com/google-research/bert. BERT has 2 models (1 is Base that has 110 million parameters and 1 is Large that has 340 million parameters) both of which are smaller than GPT-2 and GPT-3. A comparison of BERT Base and Large against the original OpenAI GPT model can be found in Table 1 of https://arxiv.org/abs/1810.04805?source=post_page. This comparison is for General Language Understanding Evaluation (GLUE) benchmark tests (see https://gluebenchmark.com/ and the original paper https://openreview.net/pdf?id=rJ4km2R5t7) and shows that BERT seems to perform better.

Comparison of BERT and GPT-3

BERT and GPT-3 use different approaches to training their models. BERT is bidirectional left to right and right to left, while GPT-3 is unidirectional left to right (see https://360digitmg.com/gpt-vs-bert). Since GPT-3 is so much larger in size, it takes less additional training data to fine tune a model with GPT-3 vs BERT. Since BERT is an older model developed by Google compare to GPT-3 which is newer and larger, I wonder if Google has a more capable model or newer version of BERT that might be able to compare better against GPT-3. I am definitely interested in a more direct comparison between BERT and GPT-3 using similar sized models to see how important bidirectional vs unidirectional training approach is. It seems like there would be more information captured in a bidirectional approach.

The paper on GPT-3 mentioned above (https://arxiv.org/pdf/2005.14165.pdf) performs comparisons of model performance across different conditions of contextual learning:

- fine tuning: updating the weights of a pre-trained model by training on a task specific dataset; typically thousands to hundreds of thousands of labeled examples are used
- few-shot: the model is given a few demonstrations of the specific task at inference time as conditioning, but no weight updates are allowed.; typically 10 to 100 demonstrations
- one-shot: same as few-shot but only one demonstration is allowed
- zero-shot: no task specific learning

The goal is to achieve zero shot performance so that context specific model tuning (even with just a few demonstrations) is not required. The research shows that the larger the model, in terms of number of parameters, the better the performance (not a surprise). Moreover, they found that few-shot performance improves faster with model size than one-shot performance improvement which is faster than zero-shot performance.

Conclusion

Transfer learning for NLP seems to be an active area of research but it appears there is still a ways to go. The goal of not needing task specific tailoring of NLP models (zero-shot) has not yet

been achieved, but that is a goal that is still viewed as possible and is continued to be pursued by researchers.